Guiding Internet-Scale Video Service Deployment Using Microblog-Based Prediction

Zhi Wang*, Lifeng Sun*, Chuan Wu†, and Shiqiang Yang*
*Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
†Department of Computer Science, The University of Hong Kong
wangzhi04@mails.tsinghua.edu.cn, sunlf@tsinghua.edu.cn, cwu@cs.hku.hk, yangshq@tsinghua.edu.cn

***Abstract*—Online microblogging has been very popular in today's Internet, where users exchange short messages and follow various contents shared by people that they are interested in. Among the variety of exchanges, video links are a representative type on a microblogging site. More and more viewers of an Internet video service are coming from microblog recommendations. It is intriguing research to explore the connections between the patterns of microblog exchanges and the popularity of videos, in order to potentially use the propagation patterns of microblogs to guide proactive service deployment of a video sharing system. Based on extensive traces from Youku and Tencent Weibo, a popular video sharing site and a favored microblogging system in China, we explore how patterns of video link propagation in the microblogging system are correlated with video popularity on the video sharing site, at different times and in different geographic regions. Using influential factors summarized from the measurement studies, we further design neural network-based learning frameworks to predict the number of potential viewers of different videos and the geographic distribution of viewers. Experiments show that our neural network-based frameworks achieve better prediction accuracy, as compared to a classical approach that relies on historical numbers of views. We also briefly discuss how proactive video service deployment can be effectively enabled by our prediction frameworks.**

I. Introduction

Recent years have seen the blossom of microblogging services in the Internet, *e.g.*, Twitter, Google+, Plurk. In a microblogging system, users can create and maintain social connections among each other, as well as publish contents or subscribe to contents shared by others from external content sharing systems, as "followers" [1]. Among the variety of contents to exchange, links to videos on video sharing sites are a popular type, and users from microblogging exchanges are constituting a large portion of the viewers in YouTube-like video sharing sites [2][3].

With increasing popularity, a microblogging system resembles the real society, and interests, beliefs, and behavior of users in such a system are representative of those in the real world [4]. Ritterman *et al.* [5] advocate to forecast a swine flu pandemic based on a belief change model summarized from Twitter. Hong *et al.* [6] and Petrovic *et al.* [7] use statistics collected from Twitter to predict popular discussion topics (messages), as well as how messages propagate among people.

A microblogging system is closely connected to many content sharing sites: it ideally samples valuable information on how users produce contents and share contents from those sites among each other, given that more and more users of a content sharing site (*e.g.*, YouTube) are coming from microblogging recommendations (*e.g.*, Twitter messages); and content sharing/propagation models in a microblogging system can be usefully exploited to improve service provision quality of a content sharing system. Nevertheless, to the best of our knowledge, this value of microblogging systems has not been investigated in the existing studies.

In this paper, we advocate to exploit the sampling and prediction capabilities of a microblogging system to provision better Internet video services. In a typical video sharing site today, large volumes of videos are uploaded by users, with viewers from all over the world. For example, more than 48 hours' worth of videos are uploaded every minute in YouTube, served to millions of users per minute. A common practice to provision these video services is to replicate videos in severs at different geographic locations [8], but it is impractical to replicate all the videos in every location. An effective, dynamic replication strategy to serve the dynamic demand for different videos from different geographic regions, is in need.

There have been proposals that build forecasting models based on historical data in a video sharing site itself, for the prediction of future views of videos [9]. We propose to exploit content sharing patterns from a microblogging system for this purpose. The potential benefits are two-folded: (1) a content sharing site typically has no information on how video views propagate among its users, while a view propagation model could enable more effective view prediction; (2) the exchanges of video links in a microblogging system typically happen earlier than the actually video views on a video sharing site, and the time lag between both events can allow more timely, proactive deployment of videos.

In our study, we explore connections between microblogging exchanges of video links and popularity of videos, based on extensive traces collected from Tencent Weibo (Weibo for short hereinafter) [10] and Youku [11]. We identify important characteristics of Weibo, which influence video access pat-

This work has been partially supported by the Development Plan of China (973) under Grant No. 2011CB302206, the National Natural Science Foundation of China under Grant No. 60833009/60933013, and the Research Grants Council of Hong Kong under Contract No. 718710E.

terns on Youku at different times and in different geographic regions: (1) the number of users that have *introduced* a video into Weibo, (2) the number of users which *re-share* links to the video to their followers, (3) how many followers that the shared video links can reach, and (4) the *geographic distribution* of Weibo users. We exploit these influential factors in the design of neural network-based learning frameworks, for predicting the number of potential viewers of different videos and the geographic distribution of viewers in a video sharing system.

The rest of this paper is organized as follows. We discuss our collection of traces in Sec. II, investigate connections between information propagation characteristics on Weibo and the number of views on Youku in Sec. III, design neural network models to predict future number and geographic distribution of views as well as evaluate the models in Sec. IV, and conclude the paper with discussions on a proactive video deployment scheme based on the prediction models in Sec. V.

II. Trace Collection and A First Glance of Connections

A. Collection of Traces

Youku is an Internet video sharing site similar to YouTube, with immense popularity in China. We crawled the number of viewers of 2291 videos from 5 most popular categories on Youku, using a crawler implemented in C# language. The videos were published between March 19 and June 20, 2011, and were the recommended ones on Youku's frontpage on June 20, 2011. The crawling was carried out on an hourly basis during June 20 to June 30, 2011. We have obtained 510,283 valid records of the view numbers for our study.

Tencent Weibo is a Twitter-like Chinese microblogging site, where users can broadcast a message including at most 140 Chinese characters to their followers. We obtained Weibo traces from the technical team of Tencent, containing valuable runtime data of the system in the whole span of June 2011. Each entry in the traces corresponds to one microblog published, including ID, name, IP address of the publisher, time stamp when the microblog was posted, IDs of the parent and root microbloggers if it is a re-post, and contents of the microblog. The traces were recorded on an hourly basis.

B. Connection between Youku and Weibo Traces

A large portion of microblogs on Weibo are sharing video links, and a dominating number of video links are URLs of videos hosted on Youku. A Weibo user who reads such a microblog may connect to Youku and watch the recommended video. We parse the traces from Weibo and only study microblog entries containing video links to Youku.

We have observed that video links on Weibo span a large portion of videos on Youku. Among the 2291 videos we investigate, 1134 were published in June 2011, and links to 93% of them (1052) appeared on Weibo in the same month. Through our study of the collected traces, we summarize the following statistics about the videos appeared on Weibo: (1) It takes as short as 172 seconds for a video to be first introduced to Weibo after its publication on Youku, and the average is 7 hours. (2) Within the first hour after a video is introduced to Weibo, as many as 527 microblogs re-share the video link, and the average number of re-sharing microblogs is 4.5 per video in the hour. (3) In this first hour, the total number of Weibo users who can see the microblogs containing the video link (*i.e.*, who are followers of the users publishing the link) is 750 on average, and 239,411 for the most popular video. These statistics show that video links exchanged on Weibo correspond to a good, timely sample set of all videos published on Youku, and it is promising to predict future video views using microblog propagation information from Weibo.

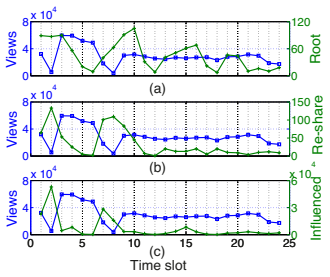


Fig. 1. Predictability of Youku video views using Weibo statistics.

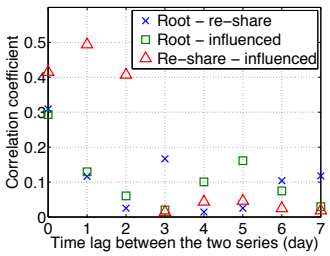


Fig. 2. Correlation among the numbers of root/re-share/influenced users.

III. Measurements and Correlation Analysis

We investigate the predictability of future Youku view numbers using historical Weibo measurements, as well as the connection between future geographic distribution of viewers and their historical distributions.

A. Predictability using Microblogging Measurements

To investigate the correlation between video link propagation on Weibo and the actual number of viewers in Youku, we study the following measurements in Weibo traces: (1) the number of users who have introduced a video into Weibo (referred to as a *root user*) by posting a microblog containing a link to the video in a time slot, (2) the number of users who after reading a microblog containing the video link, re-share the link to their followers (referred to as *re-share users*) in a time slot, and (3) the number of followers of those root and re-share users, who can see the microblogs and may potentially view the video on Youku (referred to as *influenced users*) in a time slot.

Fig. 1 plots the evolution of the number of views of a representative video on Youku, together with the evolution of the number of root/re-share/influenced users of this video on Weibo, respectively, over 100 hours after its publication. Each time slot corresponds to 4 hours. We observe that before the rise or drop of the actual view numbers on Youku, there is always a leading increase or decrease of the number of either root, re-share, or influenced users on Weibo in previous time slots. We further study the detailed correlation based on 765 videos during a 7-day span, as follows.

1) Number of Views and Number of Root Users: Different Weibo users may introduce links to the same video on Youku, each of whom becomes a root user in this video's propagation

on Weibo. The more root users a video has, the more likely the video can attract more views in the future. A correlation coefficient of 0.31 is derived between the time series of the number of root users and the time series of the number of views at the lag of 1 time slot in Fig. 1(a), *i.e.*, the number of root users at time slot T is related to the number of views at time slot $T+1$.

2) Number of Views and Number of Re-share Users: Similarly, when more users are re-sharing the links to their followers, the more views can be expected on Youku. We have computed a correlation coefficient of 0.29 between the corresponding two series in Fig. 1(b), at the lag of 1 time slot.

3) Number of Views and Number of Influenced Users: The influenced users may likely become actual viewers themselves. A correlation coefficient of 0.15 is derived between the two series in Fig. 1(c) at the lag of 1 time slot.

From the above, we can see that positive correlation exists between each Weibo measurement and the future actual number of views on Youku, but the individual correlation may not be strong enough. We seek to combine all three measurements for the prediction of future Youku view numbers. To this end, we first need to investigate if there exists any strong cross correlation among the three measurements themselves. Fig. 2 gives the computed correlation coefficients between series of each pair of the three Weibo measurements, at different time lags between the two series. We observe that the correlation between each pair of the measurements is quite weak, especially when their sampling times are further apart. Hence, these measurements can be safely used simultaneously as input features in our neural network model in Sec. IV, since the weaker the correlation among the input features is, the better the learning results are [12].

B. Geographic Distribution of Microbloggers of the Videos

For video service deployment, we need information on geographic distribution of viewers of different videos. Since Weibo users sharing a video link are "samples" of all viewers of that video on Youku, we investigate the geographic distribution of Weibo users publishing a microblog containing a link to the video, and use it to estimate the distribution of all viewers in Youku [4]. The rationale is that such microblogs are published by root and re-share users of the video, who may well have just viewed the video before posting the microblogs.

Given that the majority of viewers of Youku videos are in China [11], we consider 5 representative regions in China, namely BJ (Beijing), SH (Shanghai), SZ (Shenzhen), CD (Chengdu) and XA (Xi'an) — where large CDNs in China commonly deploy data centers [13] — and the overseas region, referred to as OS (overseas). We map the IP addresses of users in our Weibo traces to the six regions using an IP-to-location mapping database [14], and estimate the geographic distribution of viewers of video v at time T by a 6-dimensional vector $G_v^T = \{r_v^{BJ}(T), r_v^{SH}(T), r_v^{SZ}(T), r_v^{CD}(T), r_v^{XA}(T), r_v^{OS}(T)\}$, where $r_v^X(T)$ is the normalized fraction of Weibo microblogs containing links to video v, posted by users in region X in time slot T.

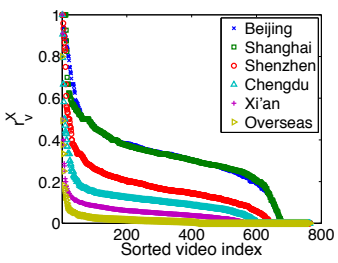


Fig. 3. Geographic distribution of Weibo users posting links to different videos.

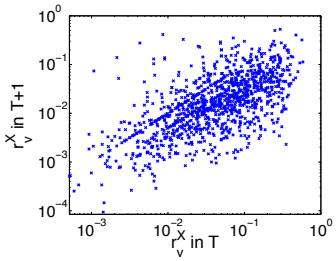


Fig. 4. Correlation between geographic distributions of microblogs in two consecutive time slots.

Skewed Geographic Distribution. Fig. 3 plots the average fraction of microblogs posted in each region containing links to each of the 765 videos, over a period of 5 days. We observe that the distribution over different regions is highly skewed: more than 40% of the viewers of the videos reside in Beijing and Shanghai regions, while very small fractions of viewers are from the overseas.

Predictability of Future Geographic Distribution. To investigate whether future geographic distribution can be predicted by historical distributions, we plot in Fig. 4 the fraction of microblogs of a video posted in a region in time slot $T+1$ versus that in the previous time slot T, for all regions and the 765 videos during a 7-day span. Each time slot is 4 hours. Strong correlation between the two can be observed, with a correlation coefficient of 0.29. We will therefore make use of historical geographic distribution to predict the future distribution, in our neural network model in Sec. IV.

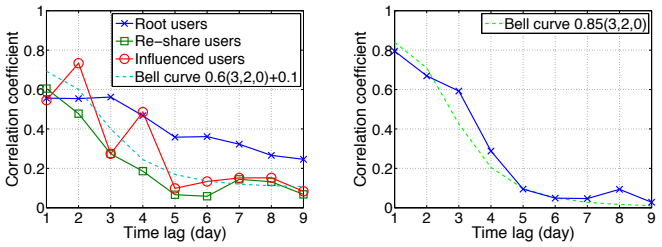
C. Influence of Measurements at Different Time Lags

Besides observing correlations between Weibo measurements at T and Youku view numbers at $T+1$, we further investigate the correlation at different time lags between the two. In Fig. 5(a), each sample represents the average correlation coefficient (over those of 100 videos) between a 10-day series of the number of views on Youku and a 10-day series of the number of root/re-share/influenced users, at different time lags between the two series. We see that the correlation weakens as the time lag becomes larger, and the correlation coefficients are quite small when the lag is larger than 7 days. Hence, we will use measurements collected in the previous 7 days for the prediction of view numbers only.

Fig. 5(b) shows correlation coefficients between the fractions of a video's microbloggers in the six regions at T and those at different time lags, where each sample is the average over the 100 videos in the 10 days. Similarly, recent geographic distributions have a larger influence on the future distribution, which will be weighted more in our prediction model in Sec. IV.

IV. Neural Network Models for View Prediction

We propose to train neural network models to predict the number and distribution of views for each video. The reason



(a) (b)

Fig. 5. Influence of Weibo measurements on the number and distribution of Youku views at different time lags.

of using neural networks lies in the following: (1) we will be dealing with unstructured features (*e.g.*, numbers of root/re-share/influenced users) with a large number of dimensions for this prediction, which can be well handled by a neural network-based learning framework [15]; and (2) neural networks have been proven effective for time series prediction [16], as what we are pursuing.

We train two neural network models, (A) one for the prediction of the total number of views of a video, and (B) the other for forecasting the geographic distribution of viewers. We use pre-labeled samples from the traces, and each sample consists of the input features and the prediction target(s). To train neural network (A), a sample is $\{X_v^T, \bar{V}_v^T\}$, where X_v^T is the set of input features and $\bar{V}_v^T$ is the vector denoting the level of view numbers in T. We classify the number of views N for a video into 5 levels: (1) $N < 500$, (2) $500 \leq N < 5000$, (3) $5000 \leq N < 10000$, (4) $10000 \leq N < 100000$, and (5) $N \geq 100000$. $\bar{V}_v^T$ is a 5-dimensional binary vector with $\bar{V}_v^T[i] = 1$ and $\bar{V}_v^T[j] = 0, \forall j \neq i$, denoting that the number of views belongs to the ith level. To train neural network (B), a sample is $\{Y_v^T, G_v^T\}$, where Y_v^T is the set of input features and G_v^T is a 6-dimensional vector, where each element represents the fraction of microblogs posted in one of the six regions (BJ, SH, SZ, CD, XA, OS) in T, respectively. We next detail how the input features are selected.

A. Selecting Input Features

We have shown the correlation between the number of Youku views in T and the number of root/re-share/influenced users in times before T, as well as the correlation between the geographic distribution of viewers in T with the historical distributions before T. To use those factors as input features, we need to decide a time window, a frequency for feature exaction in the time window (*i.e.*, the number of time slots sampled in the time window), and the weight of each feature in the learning framework.

Time Window. We have observed that the correlation between the number of Youku views and its influential Weibo measurements, as well as the correlation between the geographic distribution of viewers and its influential measurements, are weak when their time lags are larger than 7 days, respectively. Hence, we only extract features from measurements within the recent 7 days to train neural networks (A) and (B). For a newly published video with a lifetime shorter than 7 days, we use measurements throughout its past lifetime.

Frequency. We extract influential measurements as features once per day during the recent 7 days (or during its past lifetime for a new video). The features are extracted at the same time of each day, in order to capture any existing daily patterns. Let M denote the number of days the features are extracted.

Weight. Existing studies have shown that the learning performance of a neural network model can be improved by properly weighting the input features [17]. We weight the features from different time slots according to their levels of correlation with the prediction targets. In Fig. 5(a) and (b), the curves of correlation coefficients can be fitted well by generalized bell functions $f(x) = \frac{e}{1+\left|\frac{x-c}{a}\right|^{2b}} + d$ (in Fig. 5, for simplicity, we denote a particular bell function by "$e(a,b,c)+d$"). Hence, we weight the number of root/re-share/influenced users, to be used as features in neural network (A), by $\alpha(x) = \frac{0.6}{1+|x/3|^4} + 0.1$, and the past geographic distributions of viewers, which are used as features in neural network (B), by $\beta(x) = \frac{0.85}{1+|x/3|^4}$, where x is the time lag between the time slots when the prediction target and the corresponding features happen, respectively.

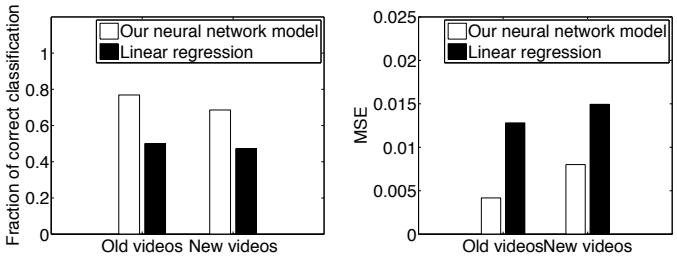
Let $R_v^T(i)$, $S_v^T(i)$, and $I_v^T(i)$ be the number of root, re-share, and influenced users of video v in the ith time slot in the time window before T, respectively. Let $G_{v,r}^T(i)$ denote the fraction of microblogs of video v posted by users in region r in the ith time slot in the time window before T. In summary, we use features $X_v^T = \{\alpha(i)R_v^T(i), \alpha(i)S_v^T(i), \alpha(i)I_v^T(i) | i = 1, 2, \ldots, M\}$ in neural network (A), and features $Y_v^T = \{\beta(i)G_{v,r}^T(i) | \forall r \in \{BJ, SH, SZ, CD, XA, OS\}, i = 1, 2, \ldots, M\}$ in neural network (B), respectively.

B. Training Neural Networks

We train a two-layer feed-forward neural network for predicting the number of views and another for forecasting geographic distribution of viewers. The training sets consist of about $29,000$ samples from the Weibo and Youku traces, corresponding to 1000 videos in 1 week's period of time (June 20 - June 26, 2011).

Neural Network (A). The *output layer* in the neural network for predicting the number of views, consists of 5 neurons, corresponding to the elements in vector $\bar{V}_v^T$ respectively. The *input layer* has $3M$ nodes, corresponding to the feature set X_v^T. In the *hidden layer*, the number of neurons is decided as follows: we vary the number of hidden neurons from 15 to 25, and measure the number of samples whose view numbers can be classified into the correct levels, using a validation set of 25% of all samples from the training set. We observe that 15 hidden neurons can achieve the best results in cases of old videos with a lifetime longer than 7 days, and 18 hidden neurons are needed for most of the new videos.

Neural Network (B). The output layer in the neural network for predicting the geographic distribution of viewers corresponds to the 6-dimensional vector G_v^T. The input layer



(a) Number of views (b) Geographic distribution of viewers

Fig. 6. Prediction accuracy: a comparison.

corresponds to the $6M$-dimensional vector, containing M vectors of viewer geographic distributions (each element is the fraction of microbloggers of video v in each of the six regions) in the previous M days. To decide the number of neurons in the hidden layer, we measure the MSE (mean squared error) between the output geographic distribution vector and the actual geographic distribution vector from the traces, by varying the number of hidden neurons from 10 to 50. Our training results give that 20 hidden neurons provide the best accuracy in cases of old videos, and 35 hidden neurons for most of the new videos.

C. Evaluating the Predication Accuracy

We evaluate the accuracy of our neural network models using 6000 samples extracted from the same set of traces where the $29,000$ training samples are extracted (but the test samples are different from the training samples). For neural network (A), we investigate the fraction of test samples whose view numbers are classified into the correct levels. For neural network (B), we evaluate the MSE between the predicted geographic distribution vector and the actual geographic distribution vector. We also compare the accuracy of our neural networks with that of a linear regression approach [18]. With the linear regression approach, a linear model is learned to predict the future number of views (geographic distribution) on Youku based on the historical numbers of views (geographic distributions) on Youku in the past M time slots, using least-square line fitting.

Fig. 6 shows the evaluation results. We observe that our neural network models achieve much better prediction accuracy than that achieved by the linear regression approach. In addition, the number of views and geographic distribution of old videos can be better forecasted, than those of the new videos, due to the fewer number of features in the learning framework of new videos.

V. Concluding Discussions

In this paper, we explore the connections between information propagation in a microblogging system and the number and distribution of actual views in a video sharing site, using extensive traces from two large-scale real-world microblogging and video sharing systems. We have made the following intriguing observations: (1) Video links exchanged in the microblogging system correspond to a timely sample set of videos on the video sharing site; (2) The number of future video views can be effectively predicted using microblogging information, *i.e.*, the numbers of root, re-share and influenced users during propagation of the video links; (3) The geographic distribution of video viewers can be sampled from the distribution of microbloggers and predicted from the historical distributions. Based on our discoveries, we develop two neural network models for predicting the future number of video views and geographic distribution of the viewers, respectively. Our cross validation experiments using large sample sets from the traces have verified the accuracy of our prediction models.

These prediction models can be effectively exploited for proactive video deployment, to improve service provision efficiency and quality. After a video is published on a video sharing site, the service provider can replicate it in servers in regions with predicted large surge of viewing demand in the near future. To implement a practical video service using our discoveries, the following issues should be addressed: (1) Microblogs have to be efficiently and dynamically collected for prediction of video views, *e.g.*, possibly via dedicated APIs provided by the microblogging system for this purpose; (2) Dynamic, timely calibration of the neural network models should be enabled with the newest input from the microblogging system. We are working on the detailed design of a proactive video service provisioning scheme based on the models, in our ongoing research.

References

[1] H. Kwak, C. Lee, H. Park, and S. Moon, "What Is Twitter, a Social Network or a News Media?" in *Proc. of ACM WWW*, 2010.
[2] K. Lai and D. Wang, "Towards Understanding the External Links of Video Sharing Sites: Measurement and Analysis," in *Proc. of ACM NOSSDAV*, 2010.
[3] "http://www.rboke.com/net/youku/2011/0714/7295.html."
[4] N. Savage, "Twitter As Medium and Message," *Communications of the ACM*, vol. 54, no. 3, pp. 18–20, 2011.
[5] J. Ritterman, M. Osborne, and E. Klein, "Using Prediction Markets and Twitter to Predict a Swine Flu Pandemic," in *Proc. of International Workshop on Mining Social Media*, 2009.
[6] L. Hong, O. Dan, and B. Davison, "Predicting Popular Messages in Twitter," in *Proc. of ACM WWW*, 2011.
[7] S. Petrovic, M. Osborne, and V. Lavrenko, "RT to Win! Predicting Message Propagation in Twitter," in *Prof. of AAAI on Weblogs and Social Media*, 2011.
[8] V. Adhikari, S. Jain, Y. Chen, and Z. Zhang, "Reverse Engineering the Youtube Video Delivery Cloud," in *Proc. of IEEE HotMD*, 2011.
[9] G. Szabo and B. Huberman, "Predicting the Popularity of Online Content," *Communications of the ACM*, vol. 53, no. 8, pp. 80–88, 2008.
[10] "http://t.qq.com."
[11] "http://www.youku.com."
[12] K. Suzuki, *Artificial Neural Networks - Methodological Advances and Biomedical Applications*. InTech, 2011.
[13] H. Yin, X. Liu, T. Zhan, V. Sekar, F. Qiu, C. Lin, H., and B. Li, "Design and Deployment of a Hybrid CDN-P2P System for Live Video Streaming: Experiences With Livesky," in *Proc. of ACM Multimedia*, 2009.
[14] "http://www.cz88.net/."
[15] L. Yu, S. Wang, and K. Lai, "An Integrated Data Preparation Scheme for Neural Network Data Analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 2, pp. 217–230, 2006.
[16] E. Azoff, *Neural Network Time Series Forecasting of Financial Markets*. John Wiley & Sons, Inc., 1994.
[17] D. Specht, "A General Regression Neural Network," *IEEE Transactions on Neural Networks*, vol. 2, no. 6, pp. 568–576, 1991.
[18] G. Marchuk, *Numerical Methods and Applications*. CRC, 1994.